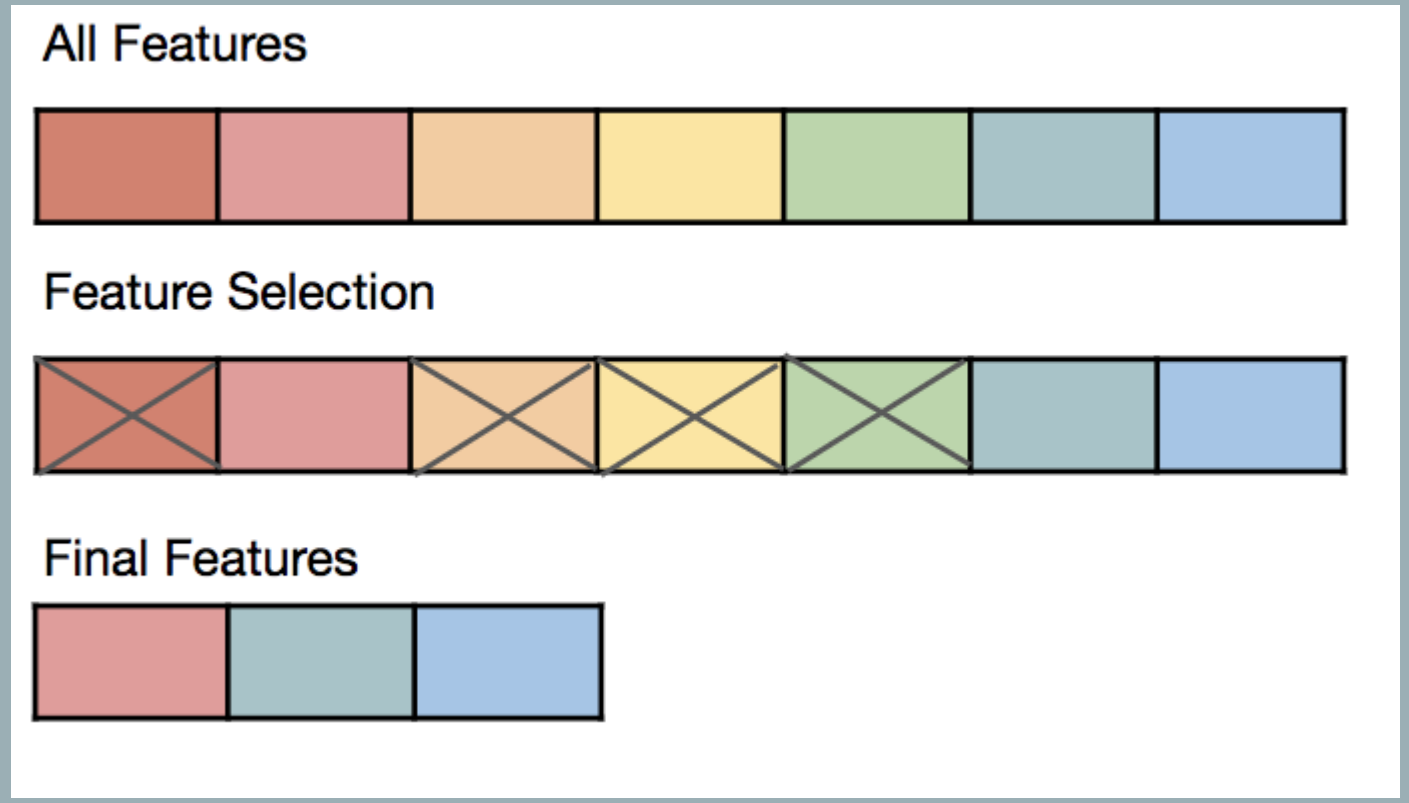
FEATURE SELECTION

It is the process of reducing the number of input variables when developing a predictive model.



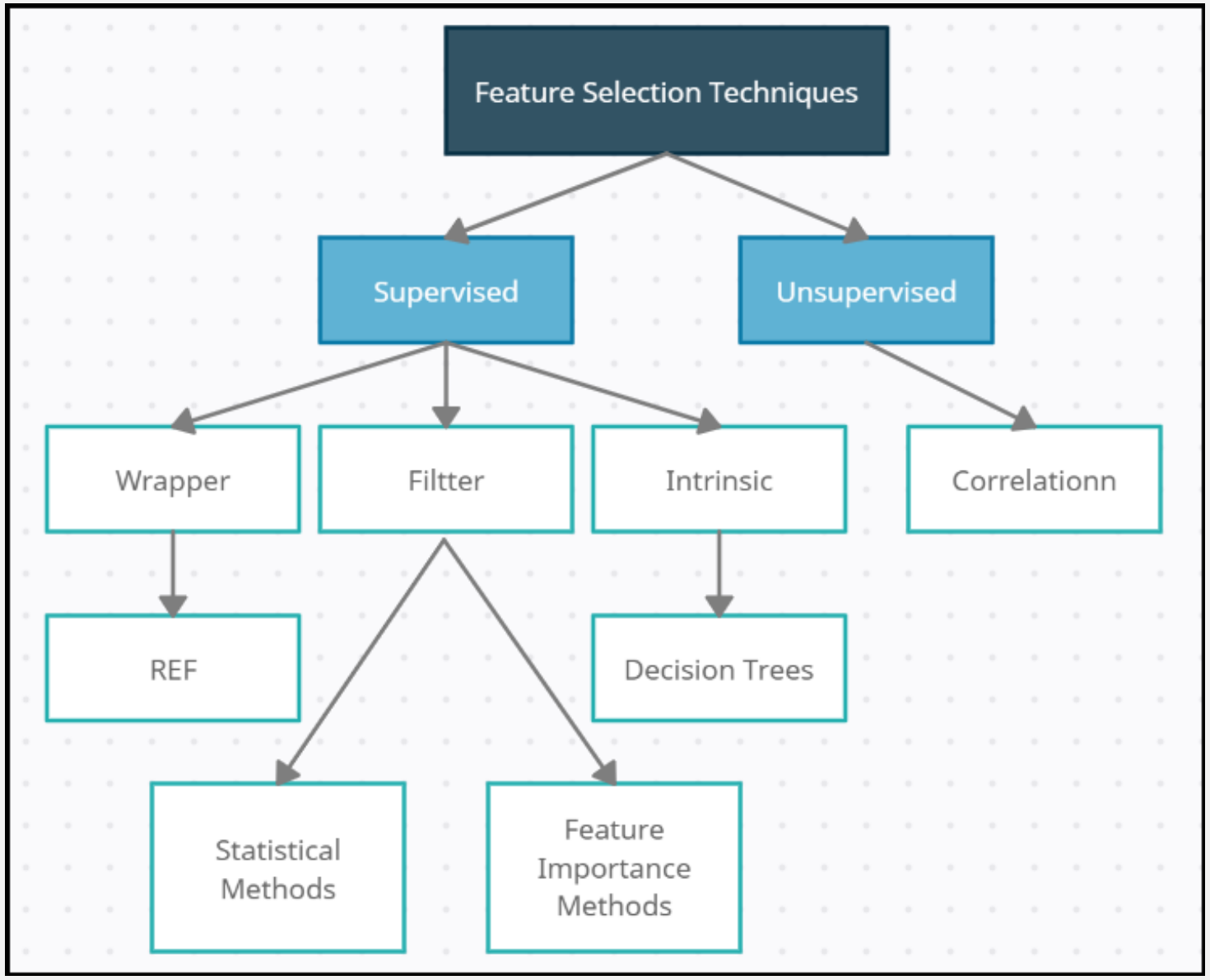
BY:
M.ASHISH REDDY
Data science with python
BATCH-5

BENEFITS OF PERFORMING FEATURE SELECTION BEFORE MODELING YOUR DATA?

1. **Reduces Overfitting**: Less redundant (*same piece of data is stored in two or more separate places*) data means less opportunity to make decisions based on noise.
2. **Improves Accuracy**: Less misleading data means modelling accuracy improves.
3. **Reduces Training Time**: fewer data points reduce algorithm complexity and algorithms train faster.

METHODS OF FEATURE SELECTION

- **Unsupervised feature selection techniques ignores the target variable**, such as methods that remove redundant variables using correlation**. Supervised feature selection techniques use the target variable**, such as methods that remove irrelevant variables..
- Another way to consider the mechanism used to select features which may be divided into **wrapper** and **filter** methods. These methods are almost always supervised and are evaluated based on the performance of a resulting model on a hold out dataset.


Feature Selection Techniques
Supervised
Unsupervised
Wrapper
Filtter
Intrinsic
Correlationn
REF
Decision Trees
Statistical Methods
Feature Importance Methods

WRAPPER FEATURE SELECTION METHOD:

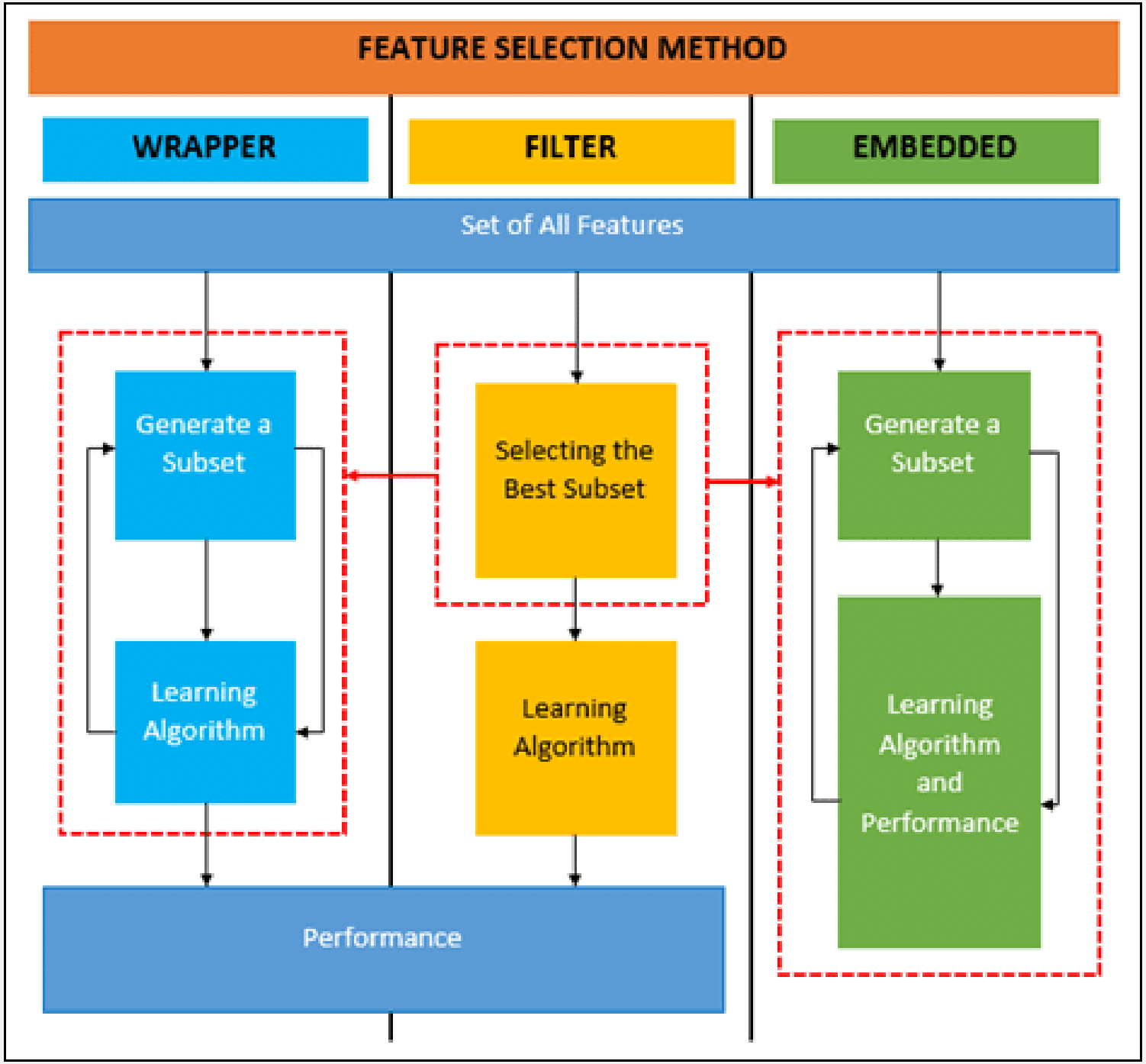
- Wrapper feature selection methods create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric. These methods are unconcerned with the variable types, although they can be computationally expensive.
- RFE(Recursive feature selection) is a good example of a wrapper feature selection method.

FILTER FEATURE SELECTION METHODS

- Filter feature selection methods use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model.

INTRINSIC FEATURE SELECTION METHODS.

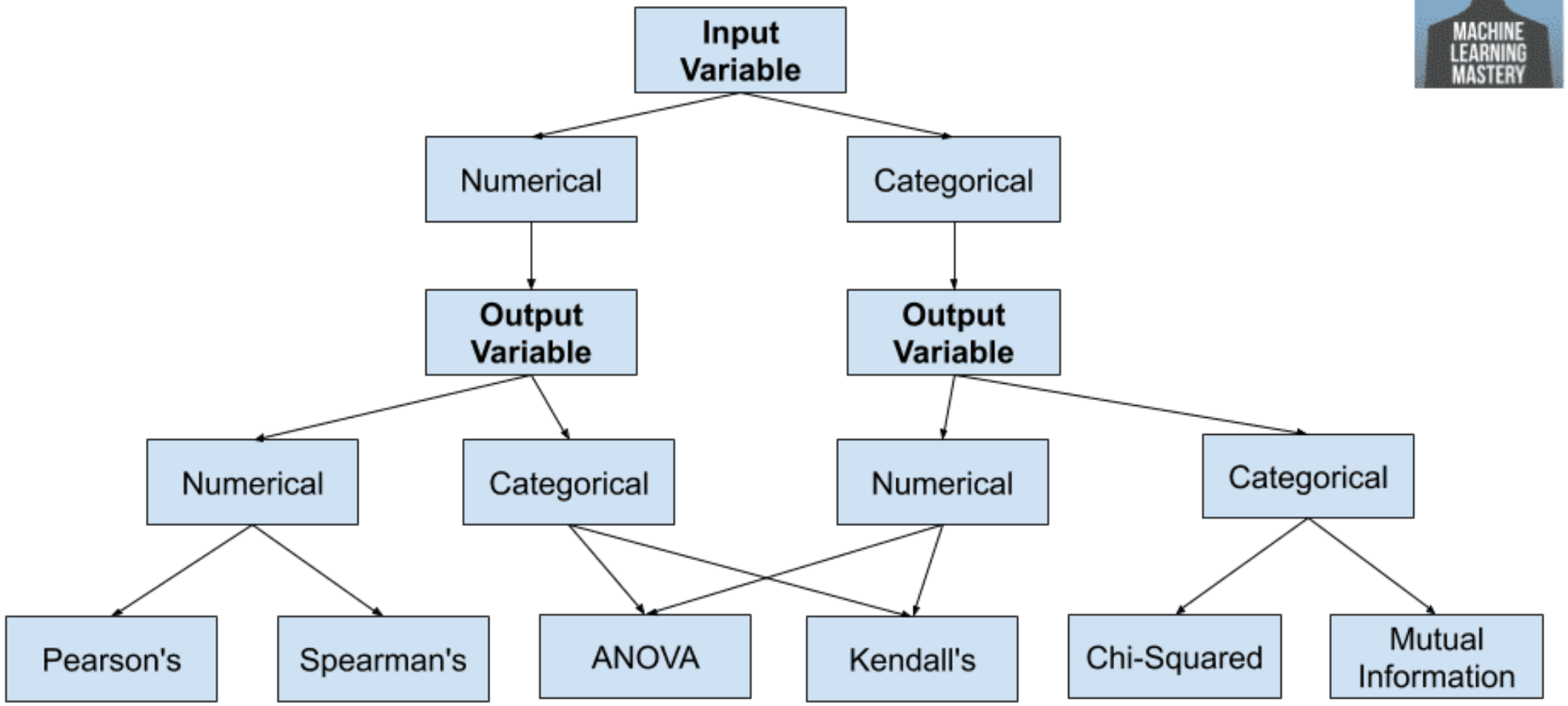
- Finally, there are some machine learning algorithms that perform feature selection automatically as part of learning the model. We might refer to these techniques as **intrinsic** feature selection methods.


FEATURE SELECTION METHOD
WRAPPER
FILTER
EMBEDDED
Set of All Features
Generate a Subset
Learning Algorithm
Selecting the Best Subset
Learning Algorithm
Generate a Subset
Learning Algorithm and Performance
Performance

SUMMARY OF FEATURE SELECTION METHODS:

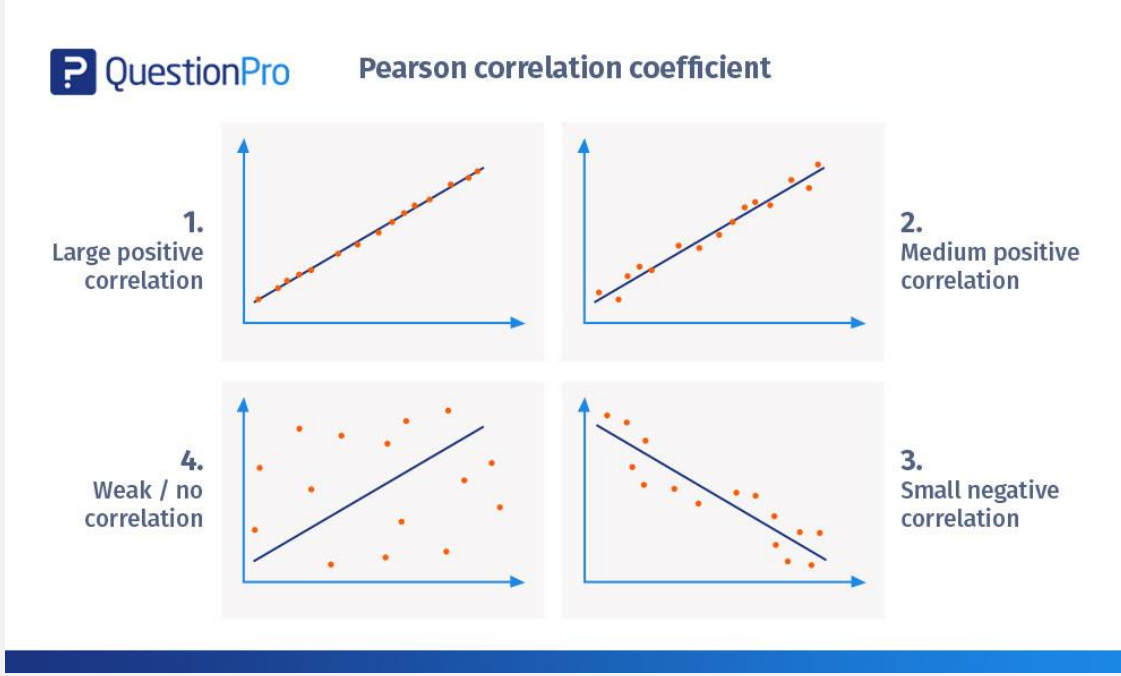
- We can summarize feature selection as follows.
- **Feature Selection**: Select a subset of input features from the dataset.
  - **Unsupervised**: Do not use the target variable (e.g. remove redundant variables).
    - Correlation
  - **Supervised**: Use the target variable (e.g. remove irrelevant variables).
    - **Wrapper**: Search for well-performing subsets of features.
      - RFE
    - **Filter**: Select subsets of features based on their relationship with the target.
      - Statistical Methods
      - Feature Importance Methods
    - **Intrinsic**: Algorithms that perform automatic feature selection during training.
      - Decision Trees
- **Dimensionality Reduction**: Project input data into a lower-dimensional feature space.

How to Choose a Feature Selection Method



NUMERICAL INPUT, NUMERICAL OUTPUT

- This is a regression predictive modeling problem with numerical input variables.
- The most common techniques are to use a correlation coefficient, such as Pearson's for a linear correlation, or rank-based methods for a nonlinear correlation.
- Pearson's correlation coefficient (linear).
- Spearman's rank coefficient (nonlinear)



NUMERICAL INPUT, CATEGORICAL OUTPUT

- This is a classification predictive modeling problem with numerical input variables.
- This might be the most common example of a classification problem,
- Again, the most common techniques are correlation based, although in this case, they must take the categorical target into account.
- ANOVA correlation coefficient (linear).
- Kendall’s rank coefficient (nonlinear).
- Kendall does assume that the categorical variable is ordinal.

CATEGORICAL INPUT, NUMERICAL OUTPUT

- This is a regression predictive modeling problem with categorical input variables.
- This is a strange example of a regression problem (e.g. you would not encounter it often).
- Nevertheless, you can use the same "*Numerical Input, Categorical Output*" methods (described above), but in reverse.

CATEGORICAL INPUT, CATEGORICAL OUTPUT

- This is a classification predictive modeling problem with categorical input variables.
- The most common correlation measure for categorical data is the chi-squared test. You can also use mutual information (information gain) from the field of information theory.
- Chi-Squared test (contingency tables).
- Mutual Information.
- In fact, mutual information is a powerful method that may prove useful for both categorical and numerical data, e.g. it is agnostic to the data types.

	Deposit	Litter	
Females	18 15	7 10	25
Males	42 45	33 30	75
	60	40	100

The formula for χ^2 is:

$$\sum \frac{(O-E)^2}{E}$$

Where O is the observed value and E is the expected value for each cell.


Thank
you!!!
...
SHUTTERSTOCK